Common Cancer Biomarkers

Christopher F. Basil,[1] Yingdong Zhao,[2] Katia Zavaglia,[1] Ping Jin,[1] Monica C. Panelli,[1] Sonia Voiculescu,[1] Susanna Mandruzzato,[4] Hueling M. Lee,[5] Barbara Seliger,[6] Ralph S. Freedman,[7] Phil R. Taylor,[3] Nan Hu,[3] Paola Zanovello,[4] Francesco M. Marincola,[1,8] and Ena Wang[1]

[1]Department of Transfusion Medicine, Warren G. Magnuson Clinical Center, [2]Biometrics Research Branch, and [3]Cancer Prevention Studies Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland; [4]Department of Oncology and Surgical Sciences, Oncology Section, University of Padua, Padua, Italy; [5]Boston Strategic Patterns, Boston, Massachusetts; [6]Institute of Medical Immunology, Martin Luther King University Hall-Wittenberg, Halle, Germany; [7]Department of Gynecology and Oncology, The University of Texas, M.D. Anderson Cancer Center, Houston, Texas; and [8]James Graham Brown Cancer Center, University of Louisville, Louisville, Kansas

Abstract

There is an increasing interest in complementing conventional histopathologic evaluation with molecular tools that could increase the sensitivity and specificity of cancer staging for diagnostic and prognostic purposes. This study strove to identify cancer-specific markers for the molecular detection of a broad range of cancer types. We used 373 archival samples inclusive of normal tissues of various lineages and benign or malignant tumors (predominantly colon, melanoma, ovarian, and esophageal cancers). All samples were processed identically and cohybridized with an identical reference RNA source to a custom-made cDNA array platform. The database was split into training (*n* = 201) and comparable prediction (*n* = 172) sets. Leave-one-out cross-validation and gene pairing analysis identified putative cancer biomarkers overexpressed by malignant lesions independent of tissue of derivation. In particular, seven gene pairs were identified with high predictive power (87%) in segregating malignant from benign lesions. Receiver operator characteristic curves based on the same genes could segregate malignant from benign tissues with 94% accuracy. The relevance of this study rests on the identification of a restricted number of biomarkers ubiquitously expressed by cancers of distinct histology. This has not been done before. These biomarkers could be used broadly to increase the sensitivity and accuracy of cancer staging and early detection of locoregional or systemic recurrence. Their selective expression by cancerous compared with paired normal tissues suggests an association with the oncogenic process resulting in stable expression during disease progression when the presently used differentiation markers are unreliable. (Cancer Res 2006; 66(6): 2953-61)

Introduction

There is an obvious interest in identifying diagnostic tools that could complement standard histopathologic evaluation to determine the presence of cancer cells in tissues (1). In particular, a pressing need exists for biomarkers useful for early cancer detection, accurate pretreatment staging, prediction of response to treatment, and monitoring of disease progression. Transcriptional profiling adds a novel dimension to cancer diagnostics identifying molecular markers capable of differentiating tumors beyond the discriminatory power of histopathologic evaluation (2). This approach not only has improved the taxonomic definition of individual cancers but also has generated novel molecular identities that could be used in association with morphologic inspection to assess the presence of cancer cells in a given tissue. A molecular approach to cancer detection is particularly useful when a few cancer cells reside in lymph nodes draining a primary site or circulate in the blood as harbingers of an impending recurrence. In those cases, although it may be difficult to identify individual cells, it may be possible to uncover the footprints of cancer through gene-specific or genome-wide amplification (3).

Most studies, however, have restricted the analysis to individual tumor types (1). This approach has limited the usefulness of the identified biomarkers for two reasons: (*a*) the biomarkers may not be broadly used as standard molecular pathology tools and (*b*) genes whose expression is irrelevant to the oncogenic process may be included. This makes current biomarkers less useful for accurate pretreatment staging, monitoring of cancer recurrence after primary treatment, and long-term follow-up of cancer patients because their expression is irrelevant to local spread, metastatization, and uncontrolled growth. Indeed, several of the currently used cancer biomarkers stem as differentiation markers from the tissue from which specific cancers originate, such as tyrosinase in melanoma (4), prostate-specific antigen in prostate cancer (5, 6), carcinoembryonic antigen in epithelial malignancies (7), and CA-125 in ovarian cancer (8). Quantitative assessment of the expression of these markers may help in the identification of cancer cells; however, their usefulness is limited by the propensity of tumor cells to progressively lose their expression (1, 9). We recently observed that the majority of genes that transcriptionally define neoplasia depend on the ontogeny of individual cancers, whereas universal oncogenic processes affect only the minority (9, 10). Therefore, like well-defined tissue differentiation markers, the expression of most genes defining a cancer histotype is likely extinguished during the natural progression of the disease (9).

The identification of cancer biomarkers related to the oncogenic process and therefore ubiquitously expressed by most malignancies could increase the sensitivity and specificity of conventional histopathologic evaluation by targeting genes whose expression is critical for invasion, metastatization, and cell survival. Transcriptional profiling of the NCI-60 cancer cell lines and a limited number

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Ena Wang, NIH, Building 10, Room 1N224B, 10 Center Drive, Bethesda, MD 20892-1184. Phone: 301-451-8501; Fax: 301-402-1360; E-mail: ewang@cc.nih.gov.

doi:10.1158/0008-5472.CAN-05-3433

of tissue specimens showed that multiclass cancer classification may lead to the identification of biomarkers expressed by different cancer types (11). Thus, the current study was aimed at the identification of common genetic traits associated with aggressiveness, uncontrolled proliferation, and metastatic potential, which could, in turn, be exploited as ubiquitous identifiers of malignancy. Therefore, we searched for genes overexpressed by cancer tissues in 373 archival cDNA microarray samples encompassing a variety of malignant and benign samples. All samples were prepared and processed identically and cohybridized consistently with a differentially labeled reference onto a 17.5K custom-made cDNA array. Novel candidate biomarkers were identified that could define malignancy with high levels of accuracy. We also tested the predictive accuracy of a list of cancer biomarkers proposed by the literature (Supplementary Data 1) and identified 332 genes included as cDNA clones in the same 17.5K array platform.

Materials and Methods

Tissue Procurement

Archival samples encompassing different tissue types (Table 1) included paired normal kidney and primary renal cell carcinoma specimens (Department of Urology, Johannes Gutenberg-University, Mainz, Germany; ref. 9); excisional biopsies of melanoma lesions (Department of Surgical Sciences, University of Padua, Padua, Italy) or fine-needle aspirates of cutaneous melanoma metastases [Surgery Branch, National Cancer Institute (NCI), NIH, Bethesda, MD; ref. 10]; primary uterine and ovarian cancers, benign ovarian lesions, and peritoneal tissues (Department of Gynecologic Oncology, M. D. Anderson Cancer Center, TX; ref. 12); primary sarcomas, one primary endometrial cancer, one primary laryngeal cancer, two primary breast cancers, and one primary colon adenocarcinoma (Tissue Network, Philadelphia, PA); primary carcinomas from the esophageal junction and paired normal esophageal tissue surrounding the tumor (Biometric Research Branch, Division of Cancer Treatment and Diagnosis, Cancer Prevention Studies Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD; ref. 2);

Table 1. Summary of samples used in this study

Training set	*n*	Prediction set	*n*
Malignant tissues			
Primary colon cancer	20	Primary colon cancer	16
Metastatic lymph nodes in patients with colon cancer	9	Metastatic lymph nodes in patients with colon cancer	9
Cutaneous melanoma metastases	8	Cutaneous melanoma metastases	3
Melanoma lymph node metastases	16	Melanoma lymph node metastases	17
In-transit metastases	2	In-transit metastases	1
Distant melanoma metastases	11	Distant melanoma metastases	10
Esophageal cancer	7	Esophageal cancer	5
Renal cell carcinoma	7	Renal cell carcinoma	7
Malignant ovarian tumor	12	Malignant ovarian tumor	14
Uterine cancer	1	Liver metastasis from colon carcinoma	1
Sarcoma	2	Sarcoma	1
Breast cancer	1	Breast cancer	1
Malignant laryngeal carcinoma	1		
Testicular cancer	1		
Total malignant lesions	98		85
Benign tissues			
Peritoneum in patient with benign ovarian pathology	6	Peritoneum in patient with benign ovarian pathology	4
Peritoneal stroma in patient with benign ovarian pathology	4	Peritoneal stroma in patient with benign ovarian pathology	4
Normal peritoneum in patients with ovarian cancer	8	Normal peritoneum in patients with ovarian cancer	6
Normal peritoneal stroma in patients with ovarian cancer	8	Normal peritoneal stroma in patients with ovarian cancer	6
Normal ovarian tissue in patient with ovarian cancer	2	Normal ovarian tissue in patient with ovarian cancer	1
Normal peritoneum in patients with uterine cancer	2	Normal peritoneum in patients with uterine cancer	1
Normal peritoneal stroma in patients with uterine cancer	2	Normal peritoneal stroma in patients with uterine cancer	3
Normal colon tissue	4	Normal colon tissue	1
Normal and hyperplastic lymph nodes in patients with colon cancer	28	Normal and hyperplastic lymph nodes in patients with colon cancer	28
Hyperplastic lymph node in normal individual	1		
Normal esophagus adjacent to cancer	6	Normal esophagus adjacent to cancer	6
Normal kidney	8	Normal kidney	7
Peripheral blood mononuclear cells	24	Peripheral blood mononuclear cells	22
Total benign lesions	103		87
Total no. lesions used for the analysis	201		172
Basal cell carcinomas	17		16
Duplicate samples	19		12
Total no. lesions included in the study	237		200

NOTE: Basal cell carcinomas and duplicate samples from the same patients were not included in any of the analyses done but were added for display in the figures.

Table 2. Summary of salient statistical analyses

Cutoff *P*	No. genes	SVM		3-Nearest neighbor		1-Nearest neighbor	
		Training (%)	Prediction (%)	Training (%)	Prediction (%)	Training (%)	Prediction (%)
Analysis based on 6,175 cDNA clones up-regulated in malignant compared with benign lesions							
$<1 \times 10^{-3}$	1516	90	ND	85	ND	89	ND
$<1 \times 10^{-7}$	395	91	ND	86	ND	86	ND
$<1 \times 10^{-13}$	50	88	ND	85	ND	88	ND
$<1.7 \times 10^{-16}$	20	90	85.5	88	86.6	82	89
$<1 \times 10^{-6}$	14 (7 pairs)	87	87	83	85	85	87
Analysis based on previously described biomarkers							
$<1 \times 10^{-6}$	50	83	88	88	90	90	88
$<5.4 \times 10^{-5}$	14 (7 pairs)	86	85	85	85	87	82

NOTE: List of genes identified applying arbitrary cutoffs of significance to differentiate benign from malignant lesions. A class prediction program was run applying incremental cutoffs of significance ($P < 1 \times 10^{-3}$, $P < 1 \times 10^{-7}$, $P < 2.8 \times 10^{-13}$, and $P < 1.7 \times 10^{-16}$) on 6,264 genes up-regulated in malignant compared with benign lesions. In addition, gene pairing was applied on the same data set eliminating genes below a significance cutoff level of $<1 \times 10^{-6}$. A similar analysis was applied to a restricted number of genes (332) that had been proposed previously as possible biomarkers (Supplementary Data 1). Salient results resulting from different statistical strategies are shown.
Abbreviation: ND, not determined.

primary colorectal carcinomas, lymph node metastases, one hepatic metastasis, and normal or hyperplastic cancer draining lymph nodes from patient with colorectal adenocarcinoma (Department of Pathology, University of Pisa, Pisa, Italy); and basal cell carcinomas (National Naval Hospital, Bethesda, MD). Specimens were collected as a result of routine operative procedures, and portions were frozen for subsequent analysis, whereas the remnant tissue was used for pathologic confirmation. Tissue procurement followed standard ethical procedure according to the institutional policy.

RNA Preparation, Amplification, and Labeling

Samples were snap frozen in the presence of RNAlater at −80°C (3). Benign samples from cancer patient were histologically evaluated for possible contamination of tumor cells. Samples with confirmed contamination were excluded.

Total RNA was extracted from frozen material using Mini or Midi kit (Qiagen, Valencia, CA) after homogenizing tissue in the presence of RLT buffer with fresh addition of 2-β-mercaptoethanol and amplified into antisense RNA (13). Although the quantity of total RNA was sufficient in most cases for gene profiling, we have shown repeatedly the high-fidelity RNA amplification yielding superior results due to lack of contaminant rRNA and tRNA (3, 13–15). Quality and quantity of total and amplified RNA were monitored using a Bioanalyzer 2000 (Agilent Technologies, Palo Alto, CA; ref. 14). Poor-quality samples were excluded. Amplified RNA from peripheral blood mononuclear cells pooled from six normal donors served as a constant reference in all experiments (3). Test and reference RNA were labeled with Cy5 (red) and Cy3 (green), respectively, and cohybridized to a custom-made 17.5K cDNA microarray printed at the Immunogenetics Section, Department of Transfusion Medicine, Warren G. Magnuson Clinical Center, Center for Cancer Research, National Cancer Institute, NIH, with a configuration of 32_24_23, and contained 17,500 elements. Clones used for printing included a combination of the Research Genetics RG_HsKG_031901 8K clone set, and 9,000 clones were selected from the RG_Hs_seq_ver_070700 40K clone set. The 17,500 spots included 12,072 uniquely named genes, 875 duplicated genes, and ∼4,000 expression sequence tags (complete gene list and printing layout are available at http://nciarray.nci.nih.gov/gal_files/index.shtml). Array quality was first validated using an internal reference concordance system based on the expectation that results obtained through the hybridization of the same test and reference material in different experiments should perfectly collimate. The level of concordance was measured by rehybridizing periodically the same arbitrarily selected test sample (A375 melanoma cell line) with the consistent reference sample as described previously (16).

Statistical Analysis

Identification of candidate biomarkers. Archival cDNA array experiments were retrieved from the NCI's microarray database eliminating those that based on image quality, background, and dye bias were considered of lower quality. The remaining 502 arrays were collated into the Biometrics Research Branch (BRB) array tool (http://linus.nci.nih.gov/BRB-ArrayTools.html) and further evaluated for quality using *M/A* plots [$M = \log_2(R/G)$, $A = \log_2\sqrt{RG}$; ref. (17)] before and after Lowess smoother normalization. Sixty-nine arrays with skewed *M/A* plots were excluded from further analysis. The remaining arrays included 33 basal cell carcinomas. These were removed from the analysis because of the ambivalent behavior of these tumors characterized by an indolent and noninvasive conduct in between malignant and benign lesions (18). Finally, only one of paired bilateral normal samples collected from the same patient (12) was used for analysis excluding additional 27 samples. Both basal cell carcinomas and paired normal samples were, however, returned to the data set for display in the figures. In the end, a total of 373 samples were used for the analysis (Table 1). These test samples were subdivided in a training set (201 arrays; 98 from malignant and 103 from benign tissues) and a prediction/validation set (172 arrays; 85 from malignant and 87 from benign samples). Class prediction comparing benign and malignant phenotypes was applied to the resulting data set using different prediction methods [compound covariant predictor, diagonal linear discriminant analysis, *k*-nearest neighbors for $k = 1$ and 3, nearest centroid, and support vector machine (SVM)] supported by the BRB array tool. Most of the information reported in this article was derived using SVM and nearest-neighbor algorithms (Table 2) that, as observed by others (19–21), outdone other approaches when applied to transcriptional profiling. Gene pair identification was based on the Greedy pairs approach (22), which starts ranking all genes based on their individual *t* scores on the training set. The procedure selects the best-ranked gene g_i and finds the other gene g_j that together with g_i provides the best discrimination using as a measure the distance between centroids of the two classes with regard to the two genes when projected to the diagonal linear discriminant axis. The two selected genes are removed from the gene set, and the procedure is repeated on the remaining set until the specified number of genes has been selected.

Class comparison was conducted using S plus program. Prediction analysis was based on leave-one-out cross-validation (LOOCV). Receiver

Table 3. Proposed biomarkers for melanoma, colon, ovarian, and esophageal carcinoma

Clone ID	UG No.	Gene	Name	First 50	First 20	Seven pairs	Refs.
781019	Hs.530077	*PON2*	Paraoxonase 2	Yes	Yes	Yes	
781019	HS.530077	*PON2*	Paraoxonase 2	Yes	Yes		
769921	Hs.93002	*UBE2C*	Ubiquitin Cong Enz E2C	Yes	Yes	Yes	(44, 53)
146882	Hs.93002	*UBE2C*	Ubiquitin Cong Enz E2C	Yes	Yes		
810928	Hs.153357	*PLOD3*	Procol-lys 1,2 oxoglute 5-dioxyg 3	Yes	Yes	Yes	
813830	Hs.289271	*CYC1*	Cytochrome *c*-1	Yes	Yes	Yes	
1613496	Hs.505172	*EST*	Unnamed	Yes	Yes	Yes	
460646	Hs.481720	*MYO10*	Myosin X	Yes	Yes	Yes	
626544	Hs.481720	*MYO10*	Myosin X	Yes	Yes	Yes	
626544	Hs.481720	*MYO10*	Myosin X	Yes	Yes		
788205	Hs.357901	*SOX4*	Sex determining region Y-box 4	Yes	Yes	Yes	
2578078	Hs.433615	*TUBB2*	Tubulin-β2	Yes	Yes	Yes	
430297	Hs.434059	*ETV4*	EST translocation variant 4	Yes	Yes	Yes	
898253	Hs.79088	*RCN2*	Reticulocalbin	Yes	Yes	Yes	
1160531	Hs.118681	*ERBB3*	V-erbB2	Yes	Yes		(37, 38)
897570	Hs.30345	*TRAP1*	Tumor necrosis factor receptor–associated protein 1	Yes	Yes		(50)
788832	Hs.515258	*GDF15*	Growth differentiation factor 15	Yes	Yes		(42)
2511265	Hs.497636	*LAMB3*	Laminin-β3	Yes	Yes		(52)
755975	Hs.76111	*DAG1*	Dystroglycan 1	Yes	Yes		(51)
2557762	Hs.458332	*PYCR1*	Pyrroline-5-carboxylate reductase 1	Yes	Yes		
2466685	Hs.471156	*ABI2*	Abl interactor 2	Yes		Yes	(34)
2565981	Hs.407995	*MIF*	Macrophage migration inhibitory factor	Yes			(41, 55)
487442	Hs.129826	*NET-5*	Tetraspanin 9	Yes			(49)
878578	Hs.513490	*ALDOA*	Aldolase A	Yes			(45)
2578793	Hs.513490	*ALDOA*	Aldolase A	Yes			(45)
897781	Hs.533782	*KRT8*	Keratin 8	Yes			(40)
824068	Hs.401903	*COX5A*	Cytochrome *c* oxidase Va	Yes			
824068	Hs.401903	*COX5A*	Cytochrome *c* oxidase Va	Yes			
2563366	Hs.518774	*PAICS*	Phosphoribosylaminoimidazole carboxylase	Yes			
139291	Hs.23616	*FLJ16517*	Unnamed	Yes			
2549991	Hs.463456	*NME2*	Nonmetastatic cells protein 2	Yes			(47)
755239	Hs.463456	*NME2*	Nonmetastatic cells protein 2	Yes			(47)
2466969	Hs.475963	*CTDSPL*	CTD small phosphatase-like	Yes			
742595	Hs.166071	*CDK5*	Cell division protein kinase 5	Yes			(43)
592802	Hs.527061	*RGS12*	Regulator of G protein signaling 12	Yes			(36)
755581	Hs.301613	*JTV1*	Unnamed	Yes			
840364	Hs.388004	*AHCY*	*S*-adenosylhomocysteine hydrolase	Yes			
290337	Hs.287412	*SVH*	Unnamed	Yes			
591864	Hs.1802	*HLADOB*	HLA class II DOβ	Yes			
2568090	Hs.512973	*HSPC121*	Butyrate-induced transcript 1	Yes			
773170	Hs.82128	*TPBG*	Trophoblast glycoprotein	Yes			
42096	Hs.491494	*CCT3*	Chaperonin containing TCP1, subunit 3	Yes			(35)
810989	Hs.9234	*NIFIE14*	Seven transmembrane domain protein	Yes			
771323	Hs.75093	*PLOD1*	Procol-lys 1,2 oxoglute 5-dioxyg 1	Yes			
823930	Hs.124126	*ARPC1A*	Actin-related protein 2/3 complex subunit 1A	Yes			
810452	Hs.517066	*TOMM34*	Translocase of outer mitochondrial membrane	Yes			
586742	Hs.524281	*WBP11*	WW domain binding protein 11	Yes			
26021	Hs.463928	*DLG4*	Unnamed	Yes			
814595	Hs.446240	*PRKCBP1*	Protein kinase C binding protein 1	Yes			(33)
739126	Hs.404119	*TSTA3*	Tissue-specific transplantation antigen P35B	Yes			
770858	HS. 374990	*CD34*	CD34			Yes	(46)
342211	HS. 467634	*OACT2*	*O*-acyltransferase domain containing 2			Yes	

operator characteristic (ROC) curves (23) were constructed by computing the sensitivity and specificity of various sets of biomarkers as discussed in Results. Principal component analysis (PCA) was applied using the Pro software program (Partek, Inc., St. Charles, MO). All *P*s are based on a two-tailed unpaired Student's *t* test. A Fisher's exact test was used to assess the significance of the classification. Data are displayed in the figures according to the central method of normalization (24).

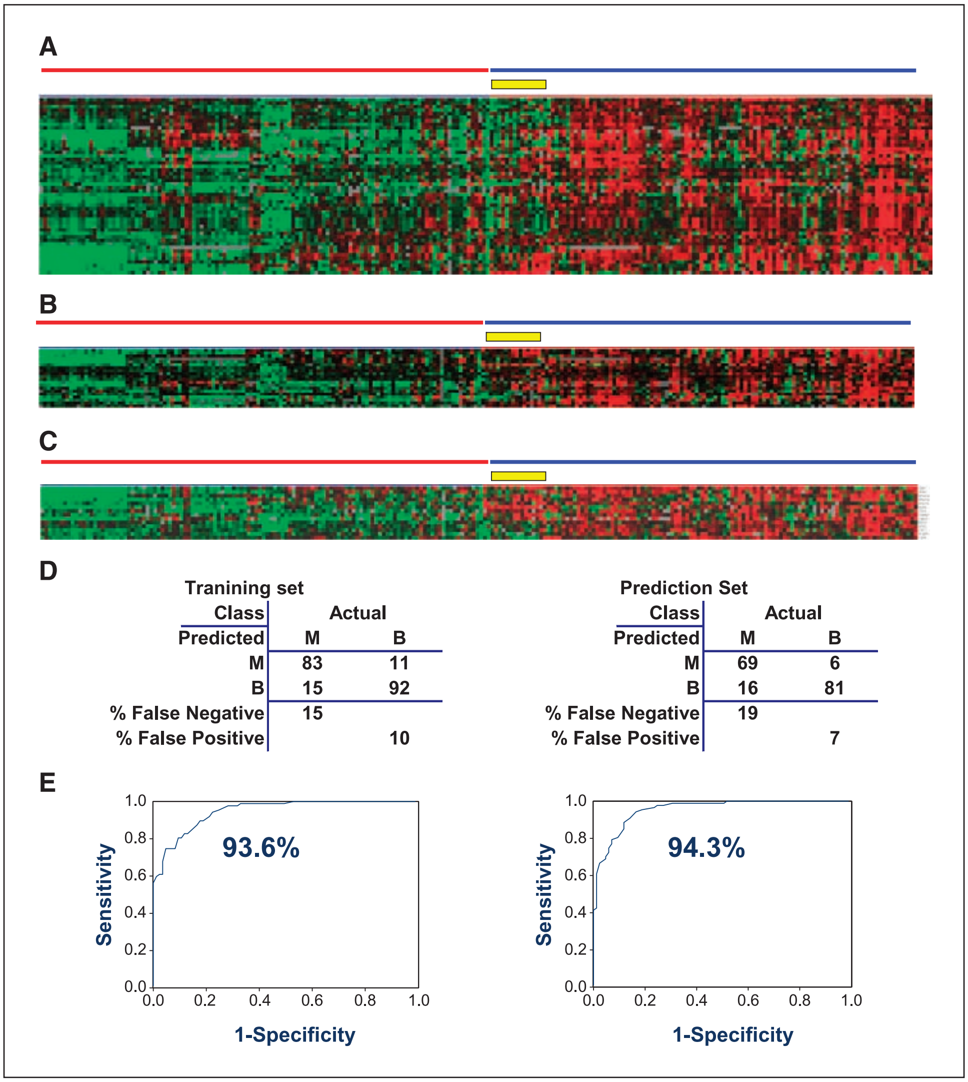
Results

Data from the arrays were filtered according to standard procedure to exclude flagged spots, spots with diameter <25 μm or intensity <200 after background subtraction. The filtered data were normalized using Lowess smoother normalization. Genes with <10% data ≤1.5-fold change in either the positive or the negative direction from the median value of the gene were excluded. In addition, genes with >20% data missing were excluded, trimming the final working set to 13,254 genes.

Scatter-plot analysis based on the average log ratio of malignant over benign lesions identified 6,264 of 13,254 genes, with a fold difference of >1 (defined as genes up-regulated in cancer). Class prediction was done by applying a univariate significance threshold ($P < 1 \times 10^{-3}$ and $P < 1 \times 10^{-7}$) to select genes suitable for LOOCV. This analysis identified 1,516 and 395 (Supplementary Data 2) genes, respectively. LOOCV based on these genes could segregate malignant from benign samples with a maximum predictive accuracy of 90% and 91%, respectively, under the SVM algorithm. The same set of genes showed lower accuracy when nearest-neighbor analysis was applied (Table 2). Individual gene expression patterns were visualized by Eisen's luster and Treeview (data not shown), showing that most genes were not exclusively expressed by malignant or benign samples, and significant overlap occurred.

To focus on the best predictors eliminating genes sporadically coexpressed by benign and malignant lesions, we ranked the 395 genes in ascending order of statistical significance (Student's *t* test; P_2 comparing malignant versus benign lesions) and selected the first 50 (Tables 2 and 3). Eisen's clustering showed a high selectivity in the pattern of expression of these genes with specific overexpression in most malignant lesions (Fig. 1*A*). LOOCV based on either SVM algorithm or 1-nearest neighbor showed a prediction accuracy of 88% when applied to the training set (Table 2). The list of genes was further narrowed to minimize the number of putative biomarkers by restricting the selection to the first 20 cDNA clones (cutoff $P < 1.7 \times 10^{-16}$; Fig. 1*B*; Tables 2 and 3) representing a total of 16 genes as some spots represented duplicates of the same gene (3 *MYO10*, 2 *PON2*, and 2 *UBE2C*). LOOCV based on the 20 cDNA clones showed a prediction accuracy of 90% in the training set.

An independent analysis was done on the complete data set of 6,264 genes up-regulated in malignant lesions by incrementally adding pairs of genes that best separated benign from malignant lesions (Greedy pairs method; ref. 22). Examination of a graphical visualization of different pairs and their corresponding accuracy yielded a logistic curve, indicating that seven gene pairs (14 genes) had prediction accuracy of 87% in the training set close to that of the 50 genes selected according to significance level (Fig. 1*C*; Tables 2 and 3). The discrimination power of the selected gene pairs was highly significant when applied to the independent prediction set



Tranining set

Class / Predicted	Actual M	Actual B
M	83	11
B	15	92
% False Negative	15	
% False Positive		10

Prediction Set

Class / Predicted	Actual M	Actual B
M	69	6
B	16	81
% False Negative	19	
% False Positive		7

Figure 1. *A,* supervised hierarchical clustering based on 50 genes ranked according to the lowest P_2s comparing benign and malignant tissue samples. Training set database: 210 samples used for the analysis plus 17 basal cell carcinomas (*yellow box*) and 19 duplicate samples obtained from the same patients that were excluded from the analysis but were used for display (Table 1). Data are according to the central method for display using a normalization factor as suggested by Ross et al. (24). Benign and malignant samples are marked with *red* and *blue lines*, respectively. *B,* 20 genes differentially expressed between malignant and benign lesions with the highest degree of significance ($P < 1.75 \times 10^{-16}$). Data as in (*A*). *C,* a seven-pair gene pairing class prediction test done on the 6,264 genes up-regulated in malignant tissue. Data as in (*A*). *D,* 2 × 2 table describing the 87% accuracy value obtained by plotting the seven gene pairs described in (*B*) in the training set (*left*) and in the prediction set (*right*) using SVM algorithm, where *B* is benign and *M* is malignant; accuracy in both case was 87%. *E,* ROC curves done using the seven gene pairs (*left*) or a set of 50 genes with the highest level of discrimination between malignant and benign samples. % Values represent the area below the curve (accuracy; ref. 23).

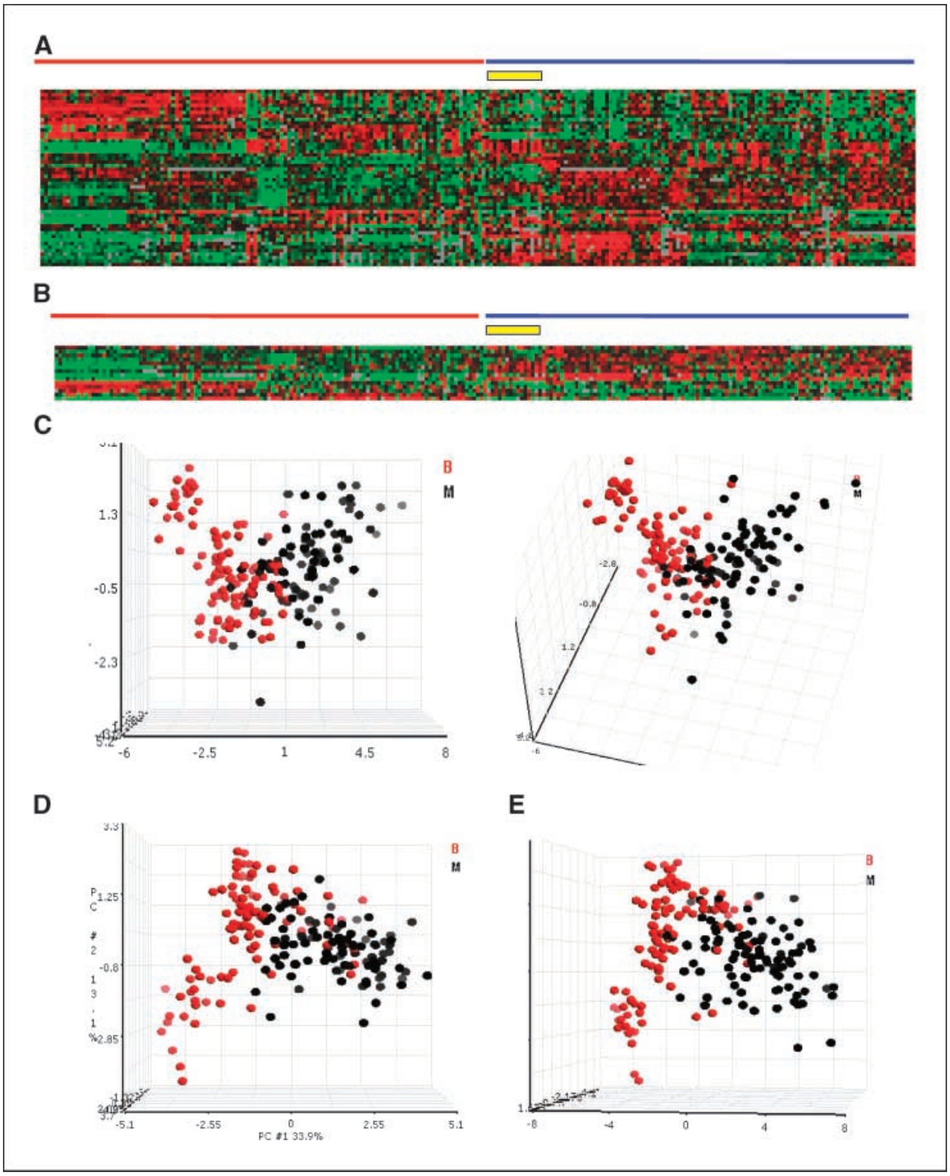


Figure 2. *A,* supervised hierarchical clustering based on 50 genes ranked according to the lowest *P*s comparing benign and malignant tissue samples. These genes were selected from a basis of 332 already proposed biomarker genes that were included in our arrays (Supplementary Data). Same training set database and additional samples described in Fig. 1*A*. Data are according to the central method for display using a normalization factor as suggested by Ross et al. (24). Benign and malignant samples are marked with *red* and *blue lines*, respectively. *B,* a seven-pair gene pairing class prediction test done on the 332 previously proposed biomarkers. Data as in (*A*). *C* to *E,* PCA based on the seven gene pairs (*red circles,* benign; *blue circles,* malignant samples) identified by analysis of the 6,264 genes up-regulated in cancer (*C*), seven gene pairs identified by the analysis of 332 known biomarkers (*D*), and using the combination of the two gene pairing results (*E*).

of 172 samples (Fisher's exact test; P <0.0001; Fig. 1*D*). Twelve of the 14 genes identified by this independent analysis were included among the 50 most significant genes, whereas 2 genes (*CD34* and *OACT2*) were not.

The predictive value of the finalist biomarkers was challenged on an independent prediction set (n = 172; Table 1). This was done in a stepwise fashion separating the prediction set into four independent groups each including ∼45 arrays. Each prediction group data was merged with the training set basing the prediction algorithm on the latter. Either the 20 most significant genes or the 14 genes identified via gene pairing (Table 2) were used to predict simultaneously the phenotype in the four separate sets. The predictive accuracy of the 14 genes for each of the four subgroups was very consistent and when combined resulted in an overall 87% maximum predictive accuracy similar to the one obtained with the training set. Interestingly, the 20-gene data set did less accurately using the SVM algorithm with a drop to 85% accuracy from the 90% observed in the training set.

Comparison of the cDNA clones and genes obtained with the two different methods (highest stringency of significance and gene pairing) showed that most of them overlapped with the exception of *AB12, CD34*, and *OACT2* that were present in the list of the 14 genes identified by gene pairing and not in the 20 most significant genes. Table 3 summarizes the various sets of genes and provides references linking their expression to cancer invasion, progression, and/or metastatization.

To further corroborate the accuracy of the biomarkers identified with the training set and confirmed with the stepwise prediction analysis, we applied direct class comparison to the combined data set (n = 379) using the S plus program. The analysis was run at a univariate cutoff $P < 1 \times 10^{-7}$ and a ratio of malignant over benign change cutoff of >2. This analysis identified 168 genes of which 11 overlapped the 14 genes identified by gene pairing. The remaining 3 genes matched the significance criteria for cutoff ($P < 1 \times 10^{-7}$) but were excluded because they were slightly below the geometric mean selected for the distinction between malignant and benign lesions.

2 × 2 Tables showed that in the majority of cases the limited accuracy was due to false-negative choices by the statistical programs, therefore decreasing overall sensitivity. This was

exemplified by the seven gene pair data set, in which false negatives occurred 15% of the times, whereas false positives occurred 10% in the training set (19% and 7% in the prediction set; Fig. 1*D*). It is of note that the majority of false-positive predictions (benign lesions predicted as malignant) occurred in samples from tissues (normal esophagus, renal epithelium, and ovarian) proximal to cancer interpreted at pathologic examination as free of cancer cell infiltration. Yet, subliminal contamination might have gone undetected. This hypothesis could not be confirmed, however, by this study because the amount of histologic material was not sufficient for further analysis. In addition, among the lesions that have been labeled as benign, there was a primary carcinoma *in situ* that was recognized as malignant by the analysis but on retrospect should have been placed *a priori* among the malignant lesions or excluded from the analysis.

The set of 14 genes was further validated using ROC curve analysis (23). This method portrays the proportion of true positives identified for any particular proportion of false positives and vice versa providing a better and more precise measure of diagnostic accuracy, because it is uninfluenced by decision biases and prior probabilities placing the performances of diverse systems on a common scale. Indeed, when ROC curves were calculated for the 14 biomarkers (Fig. 1*B*), they yielded a 93.6% accuracy underlying the superiority of this method in defining decision criteria (Fig. 1*E*). This level of accuracy was almost identical to that of the 50 most significant biomarkers (94.3%).

An extensive review of the literature and/or commercially promoted cancer biomarkers identified 332 genes present in our array platform (Supplementary Data 1). The predictive value of these genes was tested on the prediction set setting a univariate threshold $P < 0.0001$. LOOCV identified 56 genes among the 332 with significance under the set threshold. The genes were then clustered using Eisen's cluster and visualized using Treeview (Fig. 2*A*). This analysis included genes that were either up-regulated or down-regulated in malignant compared with benign lesions. Class prediction based on these genes showed a maximum 88% accuracy in correctly segregating benign and malignant samples. Gene pairing analysis done on the 332 genes identified seven gene pairs with a maximum predictive accuracy of 85% (Fig. 2*B*). Among them, only three genes (*CYC1*, *CD34*, and *ERBB3*) had also been identified by the previous analyses (Table 2). Thus, the present study identified novel ubiquitous cancer biomarkers with a prediction performance at least as good as that of the best-known cancer biomarkers. PCA showed that the best degree of separation between benign and malignant lesions could be obtained with the seven gene pairs derived analyzing the 6,264 genes overexpressed in cancer (first component score = 67.4% and second component score = 50.3%; Fig. 2*C*). PCA based on the seven gene pairs identified by analyzing the 332 known biomarkers also showed good visual separation of benign from malignant lesions (Fig. 2*D*), but the calculated discrimination was not as strong as with the first component score (56.2%) and the second component score (33.9%). The combined utilization of the 14 gene pairs did not significantly increase the discriminatory power with the first component score (56.1%) and second component score (40.1%; Fig. 2*E*).

Discussion

Global transcript analysis is a powerful taxonomic tool that can identify clinically relevant molecular subclasses of cancer otherwise not identifiable by standard pathologic examination (2). Although this discrimination has enhanced our diagnostic acumen, the identification of biomarkers whose expression is shared by most cancers could serve the general purpose of segregating malignant from benign conditions independently of individual taxonomies (11). The identified universal biomarkers could be added to the pathologist's repertoire for the uncovering of cancer invasion when comprehensive histologic evaluation is not sufficient. We have observed previously that the genetic profile of individual cancer histotypes is strongly biased toward its own ontogeny; the majority of genes preferentially expressed by each histotype represent the remnant of the cellular lineage of derivation (9, 10). Therefore, renal cell cancers share the expression of a great number of genes with normal renal epithelial cells (9) and melanomas with normal melanocytes (10, 25, 26). This tissue differentiation markers are useful when searching for ectopic cancer cells wandering in the circulation or migrating to the draining lymph nodes where, for instance, the melanoma-associated antigen tyrosinase should not be found in normal conditions (27, 28). However, the use of tissue differentiation markers for the detection of cancer is predicated on their presence and for this reason has several limitations. First, genes whose expression results from lineage differentiation have generally specialized functions not associated with cell survival in ectopic tissues. Therefore, as tumor cells migrate and dedifferentiate, their expression progressively extinguishes (29, 30). Thus, although the expression of tissue differentiation markers may be indication of cancer cell infiltration of normal tissues, lack of identification cannot exclude the presence of undifferentiated tumor cells. Second, the expression of tissue-specific markers can be affected by nonneoplastic conditions, such as demographic or behavioral factors that may decrease their specificity (5, 6, 31, 32). Finally, tissue-specific markers are normally limited to one or few cancers and, therefore, cannot be used broadly.

We compared previously the gene expression profile of normal renal epithelium with that of renal cell carcinoma tissue and cancers of other histology (9). In this three-way comparison, we recognized that a small proportion of genes were specifically overexpressed by cancers independently of the lineage derivation. We therefore extended the analysis to a larger array of tissues, including normal peripheral blood mononuclear cells as a marker of systemic infiltration of normal cells; normal hyperplastic lymph nodes draining primary colon cancer areas, which closely relate to the clinical staging of primary disease;[9] cancer-free peritoneum from patients with ovarian malignancy, which we have shown previously to harbor cancer-related signatures of inflammation that, however, are not related to the oncogenic process (12); and paired normal and cancerous epithelia (renal epithelial cells, esophageal mucosa, and normal ovary) adjacent to primary tumors judged on extensive pathologic examination to be free of cancer cells (9, 10, 16). All these tissues have been treated identically, and individual gene expression was internally controlled by a consistent reference source. We have analyzed previously the robustness, reproducibility, and concordance of this strategy comparing cDNA-based results with those obtainable with other molecular testing techniques (16).

This analysis was focused specifically on genes overexpressed by cancer tissues because these may be most useful when normal tissues are scrutinized for the presence of few, difficult to detect cancer cells. Different statistical approaches achieved rather consistent results. Of the 50 cDNA clones representative of 45

[9] K. Zavaglia et al., in preparation.

genes most significantly up-regulated in cancer using the class prediction BRB array tool, 27 were included among the 168 genes identified by direct class comparison using S plus program, whereas the remainder genes were excluded because they barely did not match the empirically set statistical thresholds. As class comparison included a variable descriptive of relative expression levels ($\log_2$ ratio $\geq$ 2 between malignant and benign tissues), the simultaneous identification of a large proportion of genes by both analyses supports not only the significance of the selection but also a substantial level of over expression in cancer tissue.

Basal cell carcinomas display a biological behavior in normal and malignant tissues with minimal local invasiveness and almost no metastatic potential (18). For this reason, these lesions were kept out of the analysis but were reintroduced in the figures to provide an intermediate biological reference (Figs. 1*A-C* and 2*A* and *B*). Visual inspection suggested that the expression pattern of most biomarkers by basal cell cancers (yellow bar) was closer to that of benign than malignant tissues, suggesting that most of the genes identified by this study are associated with aggressive behavior and metastatic potential; a conjecture also supported by the literature (refs. 11, 33–53; Table 3).

Because the purpose of the analysis was to identify a minimal number of biomarkers with the highest predictive value, we focused our interest on the 14 cDNA clones identified by gene pairing. These candidate biomarkers were validated further on the completely independent prediction set with a consistent predictive accuracy of 87%. This level of accuracy is better than the accuracy of previously reported multiclass tumor classification biomarkers identified through the analysis of cell lines (54) and challenged against a limited number of tissue samples (11, 45, 55).

The uniqueness of the current study resides in the consistency of the platform used, constant reference, strict standardization of sample processing, and stringent quality selection criteria chosen to include sample in the analysis (16). On the other hand, the usefulness of the proposed biomarkers still depends on further validation. First, the analysis compared proportional gene expression between benign and malignant tissues rather than absolute copy numbers. Thus, it is not known whether some of the genes are uniquely expressed by tumor tissues or are expressed in benign conditions although at a lower level. Second, several important tissues, particularly involving chronic or acute inflammation, were not available to us. Although we attempted to include as many relevant normal tissues as possible, further work is needed to validate the relevance of these markers in other pathophysiologic conditions. Finally, this study was done only at the transcriptional level. Thus, the proposed genes may serve, for now, as useful molecular tools to complement histopathologic examination.

Acknowledgments

Received 9/23/2005; revised 11/30/2005; accepted 1/19/2006.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Bast RC, Jr., Lilja H, Urban N, et al. Translational crossroads for biomarkers. Clin Cancer Res 2005;11:6103–8.

2. Wang E, Panelli MC, Marincola FM. Genomic analysis of cancer. Princ Pract Oncol 2003;17:1–16.

3. Wang E. RNA amplification for successful gene profiling analysis. J Transl Med 2005;3:28.

4. Quaglino P, Savoia P, Osella-Abate S, Bernengo MG. RT-PCR tyrosinase expression in the peripheral blood of melanoma patients. Expert Rev Mol Diagn 2004;4:727–41.

5. Berlin B. Prostate cancer: is the PSA test the answer? N J Med 1998;95:53–5.

6. Brawn P. Prostate-specific antigen. Semin Surg Oncol 2000;18:3–9.

7. Koness RJ. CEA: is it of value in colorectal cancer? RI Med 1995;78:164–6.

8. Markman M. Limitations to the use of the CA-125 antigen level in ovarian cancer. Curr Oncol Rep 2003;5:263–4.

9. Wang E, Lichtenfels R, Bukur J, et al. Ontogeny and oncogenesis balance the transcriptional profile of renal cell cancer. Cancer Res 2004;64:7279–87.

10. Wang E, Panelli MC, Zavaglia K, et al. Melanoma-restricted genes. J Transl Med 2004;2:34.

11. Liu JJ, Cutler G, Li W, et al. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 2005;21:2691–7.

12. Wang E, Ngalame Y, Panelli MC, et al. Peritoneal and subperitoneal stroma may facilitate regional spread of ovarian cancer. Clin Cancer Res 2005;11:113–22.

13. Wang E, Miller L, Ohnmacht GA, Liu E, Marincola FM. High-fidelity mRNA amplification for gene profiling using cDNA microarrays. Nat Biotechnol 2000;17:457–9.

14. Wang E, Marincola FM. Amplification of small quantities of mRNA for transcript analysis. In: Bowtell D, Sambrook J, editors. DNA arrays—a molecular cloning manual. 1st ed. Cold Springs Harbor (NY): Cold Spring Harbor Laboratory Press; 2002. p. 204–13.

15. Feldman AL, Costouros NG, Wang E, et al. Advantages of mRNA amplification for microarray analysis. Biotechniques 2002;33:906–14.

16. Jin P, Zhao Y, Ngalame Y, et al. Selection and validation of endogenous reference genes using a high-throughput approach. BMC Genomics 2004;5:55.

17. Yang YH, Dudoit S, Luu P, et al. Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.

18. Wong CS, Strange RC, Lear JT. Basal cell carcinoma. BMJ 2003;327:794–8.

19. Brown MP, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 2000;97:262–7.

20. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906–14.

21. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98:15149–54.

22. Bo T, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biol 2002;3:RESEARCH0017.

23. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285–93.

24. Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24:227–35.

25. Wang E, Miller LD, Ohnmacht GA, et al. Prospective molecular profiling of subcutaneous melanoma metastases suggests classifiers of immune responsiveness. Cancer Res 2002;62:3581–6.

26. Marincola FM, Wang E, Herlyn M, Seliger B, Ferrone S. Tumors as elusive targets of T cell-directed immunotherapy. Trends Immunol 2003;24:334–41.

27. Kammula US, Ghossein R, Bhattacharya S, Coit DG. Serial follow-up and the prognostic significance of reverse transcriptase-polymerase chain reaction-staged sentinel lymph nodes from melanoma patients. J Clin Oncol 2004;22:3989–96.

28. Wascher RA. Pitfalls in the use of RT-PCR as a prognostic indicator in melanoma. J Clin Oncol 2005;23:3863–4.

29. Marincola FM, Jaffe EM, Hicklin DJ, Ferrone S. Escape of human solid tumors from T cell recognition: molecular mechanisms and functional significance. Adv Immunol 2000;74:181–273.

30. Ohnmacht GA, Wang E, Mocellin S, et al. Short term kinetics of tumor antigen expression in response to vaccination. J Immunol 2001;167:1809–20.

31. Meigs JB, Mohr B, Barry MJ, Collins MM, McKinlay JB. Risk factors for clinical benign prostatic hyperplasia in a community-based population of healthy aging men. J Clin Epidemiol 2001;54:935–44.

32. Gray MA, Delahunt B, Fowles JR, Weinstein P, Cookes RR, Nacey JN. Demographic and clinical factors as determinants of serum levels of prostate specific antigen and its derivatives. Anticancer Res 2004;24:2069–72.

33. Fabbro D, Kung W, Roos W, Regazzi R, Eppenberger U. Epidermal growth factor binding and protein kinase C activities in human breast cancer cell lines: possible quantitative relationship. Cancer Res 1986;46:2720–5.

34. Shibuya N, Taki T, Mugishima H, et al. t(10;11)-acute leukemias with MLL-AF10 and MLL-ABI1 chimeric transcripts: specific expression patterns of ABI1 gene in leukemia and solid tumor cell lines. Genes Chromosomes Cancer 2001;32:1–10.

35. Midorikawa Y, Tsutsumi S, Taniguchi H, et al. Identification of genes associated with dedifferentiation of hepatocellular carcinoma with expression profiling analysis. Jpn J Cancer Res 2002;93:636–43.

36. Rae FK, Stephenson SA, Nicol DL, Clements JA. Novel association of a diverse range of genes with renal cell carcinoma as identified by differential display. Int J Cancer 2000;88:726–32.

37. Zhou H, Liu L, Lee K, et al. Lung tumorigenesis associated with erb-B-2 and erb-B-3 overexpression in human erb-B-3 transgenic mice is enhanced by methylnitrosourea. Oncogene 2002;21:8732–40.

38. Kobayashi M, Iwamatsu A, Shinohara-Kanda A, Ihara S, Fukui Y. Activation of ErbB3-3-kinase pathway is correlated with malignant phenotypes of adenocarcinomas. Oncogene 2003;22:1294–301.

39. Boyd RS, Adam PJ, Patel S, et al. Proteomic analysis of the cell-surface membrane in chronic lymphocytic leukemia: identification of two novel proteins, BCNP1 and MIG2B. Leukemia 2003;17:1605–12.

40. Casanova ML, Bravo A, Martinez-Palacio J, et al. Epidermal abnormalities and increased malignancy of skin tumors in human epidermal keratin 8-expressing transgenic mice. FASEB J 2004;18:1556–8.

41. Hagemann T, Wilson J, Kulbe H, et al. Macrophages induce invasiveness of epithelial cancer cells via NF-κB and JNK. J Immunol 2005;175:1197–205.

42. Cheung PK, Woolcock B, Adomat H, et al. Protein profiling of microdissected prostate tissue links growth differentiation factor 15 to prostate carcinogenesis. Cancer Res 2004;64:5929–33.

43. Yim EK, Meoyng J, Namakoong SE, Um SJ, Park JS. Genomic and proteomic expression patterns in HPV-16 E6 gene transfected stable human carcinoma cell lines. DNA Cell Biol 2004;23:826–35.

44. Dairkee SH, Ji Y, Ben Y, Moore DH, Meng Z, Jeffrey SS. A molecular "signature" of primary breast cancer cultures; patterns resembling tumor tissue. BMC Genomics 2004;5:47.

45. Tomonaga T, Matsushita K, Yamaguchi S, et al. Identification of altered protein expression and post-translational modifications in primary colorectal cancer by using agarose two-dimensional gel electrophoresis. Clin Cancer Res 2004;10:2007–14.

46. El Kenawy AE, Lotfy M, El Kott A, El Shahat M. Significance of matrix metalloproteinase 9 and CD34 expressions in esophageal carcinoma: correlation with DNA content. J Clin Gastroenterol 2005;39:791–4.

47. Kidd EA, Yu J, Li X, Shannon WD, Watson MA, McLeod HL. Variance in the expression of 5-fluorouracil pathway genes in colorectal cancer. Clin Cancer Res 2005;11:2612–9.

48. Haass NK, Smalley KS, Li L, Herlyn M. Adhesion, migration, and communication in melanocytes and melanoma. Pigment Cell Res 2005;18:150–9.

49. Hong IK, Kim YM, Jeoung DI, Kim KC, Lee H. Tetraspanin CD9 induces MMP-2 expression by activating p38 MAPK, JNK, and c-Jun pathways in human melanoma cells. Exp Mol Med 2005;37:230–9.

50. Macleod K, Mullen P, Sewell J, et al. Altered ErbB receptor signaling and gene expression in cisplatin-resistant ovarian cancer. Cancer Res 2005;65:6789–800.

51. Fletcher GC, Patel S, Tyson K, et al. hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4.4a and dystroglycan. Br J Cancer 2003;88:579–85.

52. Gontero P, Banisadr S, Frea B, Brausi M. Metastasis markers in bladder cancer: a review of the literature and clinical considerations. Eur Urol 2004;46:296–311.

53. Israeli O, Goldring-Aviram A, Rienstein S, et al. *In silico* chromosomal clustering of genes displaying altered expression patterns in ovarian cancer. Cancer Genet Cytogenet 2005;160:35–42.

54. Ooi CH, Tan P. Genetic algorithms applied to multiclass prediction for the analysis of gene expression data. Bioinformatics 2003;19:37–44.

55. Welsh JB, Sapinoso LM, Kern SG, et al. Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. Proc Natl Acad Sci U S A 2003;100:3410–5.